Published in: Mahlberg, Michaela/González-Díaz, Victorina/Smith, Catherine (eds.): Proceedings of the Corpus Linguistics Conference 2009, Liverpool. - Liverpool: University of Liverpool, 2010. 19 Pp.

Colligation patterns in a corpus and their lexicographic documentation

Petra Storjohann
Institut für Deutsche Sprache Mannheim
R5, 6-13
68163 Mannheim
Germany
storjohann@ids-mannheim.de

Abstract
This paper shows how corpora and related tools can be used to analyse and present significant colligational patterns lexicographically. In German, patterns such as *das nötige Wissen vermitteln* and *sein Wissen unter Beweis stellen* play a vital role when learning the language, as they exhibit relevant idiomatic usage and lexical and syntactic rules of combination. Each item has specific semantic and grammatical functions and particular preferences with respect to position and distribution. An analysis of adjectives, for example, identifies preferences in adverbial, attributive, or predicative functions.

Traditionally, corpus analyses of syntagmatic constructions have not been conducted for lexicographic purposes. This paper shows how to utilise corpora to extract and examine typical syntagms and how the results of such an analysis are documented systematically in ELEXIKO, a large-scale corpus-based Internet reference work of German. It also demonstrates how this dictionary accounts for the lexical and grammatical interplay between units in a syntagm and how authentic corpus material and complementary prose-style usage notes are a useful guide to text production or reception.

1 Colligation patterns

Combinational patterns such *das nötige Wissen vermitteln (to impart necessary knowledge to sb), sein Wissen unter Beweis stellen (to prove one's own knowledge to sbd*) and *das erworbene Wissen in die Praxis umsetzen (to put acquired knowledge into practice)* are of interest to the German linguist because like no other pattern they illustrate the characteristics of the lexico-syntactic interface of a lexical unit in its contextual environment. They are complex syntagms with lexico-syntactic relations which hold between lexical units and co-occur as typical contextual collocates (cf. Sinclair 1996, Hoey 2005). This phenomenon is referred to as 'colligation'. Colligation patterns are prototypical and regular phrasal patterns which co-occur habitually on the basis of co-selection. The interface between lexical and grammatical functions within such patterns is stressed by Tognini-Bonelli as follows:

> Co-selection will involve the co-occurrence of lexical items, identified in the concordance as strong collocational patterns, and the co-occurrence of grammatical patterns which will be identified in the concordance as colligational features, a mixture between the two almost inevitably bringing together lexical and grammatical patterns. (Tognini-Bonelli 2001: 101)

In colligational patterns, each item has specific semantic and grammatical functions, and particular preferences with respect to position and distribution in a syntagmatic sequence. Compared to collocation structures, where specific lexical but not syntactic affinities for co-occurrence are relevant, colligation refers to both lexical and syntactic cohesion that exists between items in a syntagmatic string. With reference to English, for example, Stubbs (2001: 65) describes colligation as

> the relation between a pair of grammatical categories or, in a slightly wider sense, a pairing of lexis and grammar. For example, the word form *cases* frequently co-occurs with the grammatical category of quantifier, in phrases such as *in some cases, in many cases*.

The analysis of such usage patterns that are apparent in the environment of a lexical item reveals that the subject of colligation is, of course, closely connected to questions of lexical and syntactic, as well as semantic and pragmatic, constraints. These are, for example, the choices of

the definite article in a specific position or certain prepositional preferences, all of which can be identified through the analysis of more complex habitual syntagmatic units. It is believed that prototypical syntagmatic clusters are habitually used in very similar lexical-semantic contexts exhibiting the same grammatical surroundings, and that they are part of similar textual environments, possibly used within the same genre and social context (cf. Hoey 2005 and 2009). Semanticists have tried to account for such features in theories of language, as illustrated by Hoey's Theory of Lexical Priming (2005) and Hunston & Francis's Pattern Grammar (2000). In more practical works, such as the work of lexicographers, and here of German lexicographers in particular, there is a need to develop a greater interest in presenting colligation structures much more effectively.

2 Syntagmatic patterns in foreign language acquisition

Why is information on colligation important? The above-mentioned characteristics, such as lexico-grammatical affinities, and the similar textual and social contexts of colligational patterns, might be 'primed' (cf. Hoey 2005) to be expected by native speakers. Speakers of a language community are exposed to natural discourse where a word becomes "cumulatively loaded with the contexts and co-texts in which it is encountered" (Hoey 2005: 8). As a result, native speakers acquire knowledge about which words habitually co-occur and in what kinds of contexts. The same holds for phrases and larger constructions which are recurrently used in situations of actual language use. For native speakers, this acquired knowledge becomes internalised, implicit knowledge of how adjacent lexical choices affect each other. However, learners of a foreign language (in this case German) need specific documentation and explanations as to how to recognise and reproduce such language patterns. As Hanks (1987: 121) points out, information on colligation structures are important for dictionary users and for non-native speakers in particular. Such patterns play a vital role when learning a foreign language because they exhibit typical contextual usage and lexical and syntactic rules of combination. With regard to English, and this also holds true for German, Hanks (1987: 121) comments:

> This information is, of course, of great value to dictionary users. A user who is attempting to decode text is more likely to encounter the word in one of these common, typical patterns and structures than in other possible but rare structures, while a foreign learner who is struggling to encode English naturally and idiomatically needs guidance precisely on what is typical rather than on what is possible. There is therefore the strongest possible motivation for lexicographers to spell out these facts.

Typical syntagmatic patterns provide information on the habitual usage of lexical items in their immediate context. An analysis of adjectives, for example, shows whether they exhibit preferences for adverbial, attributive, or predicative functions. The investigation of the colligational behaviour of a noun, on the other hand, demonstrates the typical distribution of subject or complement functions or shows whether it is typically used with a determiner, together with modifiers or in the plural form. And for a verb, it is interesting to see whether it is more frequently attested in its infinitive form or with respect to a specific tense form. This type of information needs to be learned when acquiring a foreign language for the purpose of text decoding and production.

3 Lexicographic documentation

It is almost a truism in linguistic theory that lexical-semantic and syntactic phenomena are deeply interwoven and that any approach to language has to accommodate both of these aspects. For the field of lexicography, however, it remains a desideratum to provide adequate descriptions of the lexico-syntactic structures of a head word and to present those structures appropriately. With regard to English lexicography, Busse & Schröder (2009) point out that it

seems impossible to integrate adequately questions of lexis and grammar into a single reference work. This seems to be even more true with respect to German lexicography. There are two main reasons why German lexicography in particular has struggled to implement recent linguistic and metalexicographic demands.

Unlike in the UK, there is no tradition in Germany of cooperation between lexicographers in academia and those in publishing companies. The craft of dictionary-making largely lies in the hands of publishing houses, whose objective is not necessarily to compile radically different dictionaries or to reflect new linguistic perspectives on language in their works, but rather to publish reference books which can be sold. And while other European lexicographic projects started to use corpora and computer-assisted analysing and compiling procedures at the very beginning of corpus linguistics (e.g. COBUILD 1987), the first attempts to conceptualise a corpus-based dictionary of German date back only to 1997 to the ELEXIKO project at the *Institut für Deutsche Sprache* (cf. Haß 2005: 13). It was only in 2003 that the first corpus-based dictionary entries were published. Until then, and this is still largely the case, lexicographic enterprises in publishing companies hardly made any use of comprehensive corpora, and they did not "exploit" their electronic database empirically to gain new information. Instead, they used it to gain quick and comfortable access to citations or indications of frequency, or to update lists of headwords for new editions. There is a huge gulf between linguistic and lexicographic endeavours in academia on the one hand, and the standard mass-produced dictionaries published by large publishing houses on the other. As a result, there is still no general corpus-based lexicographic standard in Germany, a situation which is not much different from the previous lexicographic oligopoly in Germany identified and discussed critically by Teubert (1998: 145ff).

3.1 Dictionaries for learners of German as a foreign language

Wissen (*knowledge*), a lexical item rich in colligational contexts, is taken here as an example, and its lexicographic entries in two German reference works are examined critically below. The dictionaries under examination are the two pedagogical reference books, LANGENSCHEIDT and WAHRIG GROßWÖRTERBUCH, both of which are specifically designed for learners of German. The entries given there contain the following information.

Wis|sen <das;-s; unz.> **1** *Kenntnisse, Gelehrsamkeit* **2** *Kenntnis, Bewusstsein (von etwas)* ► meines ~s <Abk.: m. W.>, unseres ~s <Abk.: u. W.> ist er schon abgereist *soweit mir, uns bekannt ist*; er besitzt, hat ein umfangreiches ~ ► mit jmd. ~; eine Aussage nach bestem ~ und Gewissen machen *ehrlich u. so gut man es weiß*; es is ohne mein ~ geschehen *ohne dass ich es wusste*; ich habe es wider besseren ~ getan *ich habe es getan, obwohl ich wusste, dass es falsch war*

(Dictionary entry *Wissen* taken from LANGENSCHEIDT (1999) Großwörterbuch Deutsch als Fremdsprache.)

Wis·sen *das*; *-s*; *nur Sg* **1. das W. in etw**. (*Dat*) die Gesamtheit der Kenntnisse (auf e-m bestimmten Gebiet od. überhaupt) <enormes, großes, umfassendes W.; sich W. aneignen; sein Wissen in Biologie, Mathematik *usw*> || **K-:** ***Wissens-, -gebiet*** || ***-K: Grund,- Schul-, Spezial-*** **2**. das W. über etw. (*Akk*)|*gespr auch* **von etw.** die Kenntnis e-r bestimmten Tatsache, e-s bestimmten Sachverhalts *o.Ä.*: *Sein W. über die Zusammenhänge in diesem Fall ist von großer Bedeutung* **3. *das W. um etw***. *geschr*; die bewusste Kenntnis e-s Sachverhalts; *Trotz seines Wissens um die Brisanz der Sache hat er mit der Presse geredet* || **ID *W. ist Macht*** wer viel weiß, kann über andere Macht ausüben; ***meines (unseres) Wissens*** *Abk* ***m. W. (u. W.)*** soviel ich weiß (soviel wir wissen); ***etw. gegen / wider sein besseres W. tun*** etw. tun, obwohl man sich bewusst ist, dass es falsch od. unrecht ist; etw. ***nach bestem Wissen und Gewissen tun*** etw. voll bewusst u. in voller Verantwortung tun; ***ohne j-s W.*** ohne dass j-d davon weiß

(Dictionary entry *Wissen* taken from WAHRIG (2008) Großwörterbuch Deutsch als Fremdsprache.)

Neither entry provides dictionary users with convincing documentation and explanations of more complex phrases and patterns, nor do they offer a systematic and user-friendly presentation of syntagmatic structures and their possible optional variants. In the case of the LANGENSCHEIDT dictionary, users would also struggle to allocate the given phrases to a context, since the pattern details are not given for a specific sense.

It is now commonly agreed that patterns derive from regular language usage as, for example, observable in a corpus. And the methodological opportunities presented by these new approaches to our understanding of a word's meaning, of language in use and of the production of phrasal constructions show that the lexicographer's craft needs to be rethought. As Busse & Schröder (2009) note, linguistic description in reference works tends today to be based on corpus evidence. Although the study of corpora is now commonplace and colligational patterns are subject to concrete observation, most German dictionaries share a methodology such that the phenomena are postulated and sometimes tested against corpus data, but they are not usually retrieved, observed and analysed using computer-held corpora. For various practical reasons, German dictionaries often continue to lack empirical approaches to the description of language use, as a result of which these two German learners' dictionaries are not valuable usage handbooks. The results are entries which do not reflect current usage and which do not contain examples of real language in actual communicative situations. They disregard empirical evidence and present possible, instead of typical, usage in an inadequate way.

3.2 ELEXIKO – a corpus-based dictionary of contemporary German

ELEXIKO is an online dictionary with the primary objective of documenting reliable information on contemporary German language usage.[1] It is a comprehensive dictionary which is currently being compiled on the basis of a large corpus at the Institut für Deutsche Sprache in Mannheim. This reference work aims at providing descriptions of words and phrases in use and at documenting the lexical, semantic and grammatical preferences and options for a word in context and discourse. Although it is not explicitly designed for learners of German, much of the information which is presented in various sections is particularly relevant for non-native users. Information on typical patterned sequences is one example where the needs of language learners have been given special attention (cf. also Storjohann 2005, Storjohann et al. 2007). These are documented in the section headed "Typische Verwendungen", a part of the dictionary which refers solely to a specific single sense of a lexical item, which means that the patterns are assigned to their corresponding semantic interpretation (see figure 1).

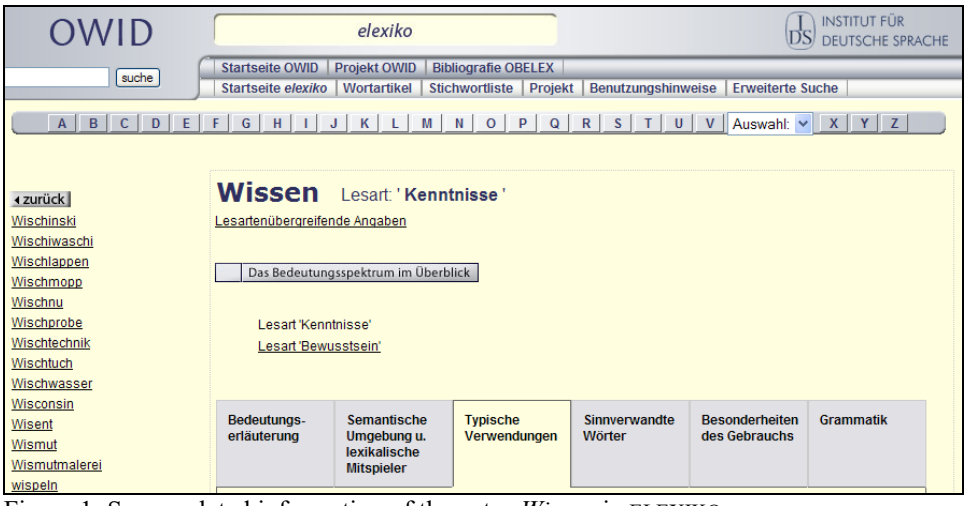

Figure 1: Sense-related information of the entry *Wissen* in ELEXIKO.

Importantly, the patterns documented here exclude true idioms, phrases or fixed expressions whose meanings cannot be derived from the elements of the lexical sequence. It is multi-word units which are recorded here, where each constitutive item can be semantically and syntactically identified, and the meaning as a whole is transparent. It is not claimed here that the two types of phrases can be clearly distinguished. On the contrary, from a theoretical point of view, all patterns should be subsumed into a general theory of lexico-syntactic construction. However, taking a lexicographic stance, some decisions (e.g. disambiguation) have to follow stricter categorisation for more practical reasons.

Typical collocational patterns, on the other hand, are documented in the neighbouring section "Semantische Umgebung und lexikalische Mitspieler" (semantic context and lexical co-occurrences). It primarily lists simple binary collocational preferences, i.e. lexical-semantic co-occurrence (e.g. adjective + noun such as *aktuelles Wissen, fundiertes Wissen, vorhandenes Wissen,* and noun + verbs collocations such as *Wissen abfragen, Wissen weitergeben, Wissen umsetzen*) rather than more complex lexico-grammatical patterns. Although separated lexicographically, both kinds of information are closely linked with each other lexicologically.

3.2.1 How data is presented

ELEXIKO bases its information about significant combinational patterns on a model which accounts for the lexical and grammatical interplay between units in a syntagmatic string, and its main objective is to present both the semantic and syntactic functions of a head word in use. The following examples serve to illustrate that ELEXIKO follows a method of presentation which underlies a systematic classification, one which groups similar patterns together, which extensively illustrates prototypical methods of combination and which presents patterns in their corresponding semantic senses.

Nouns in context, such as *Wissen* (see figure 1), are typically embedded into patterns with adjacent adjectival attributes, into verbal phrases and sentences, into patterns where they function as attributes and into sequences where the noun is followed by a preposition.

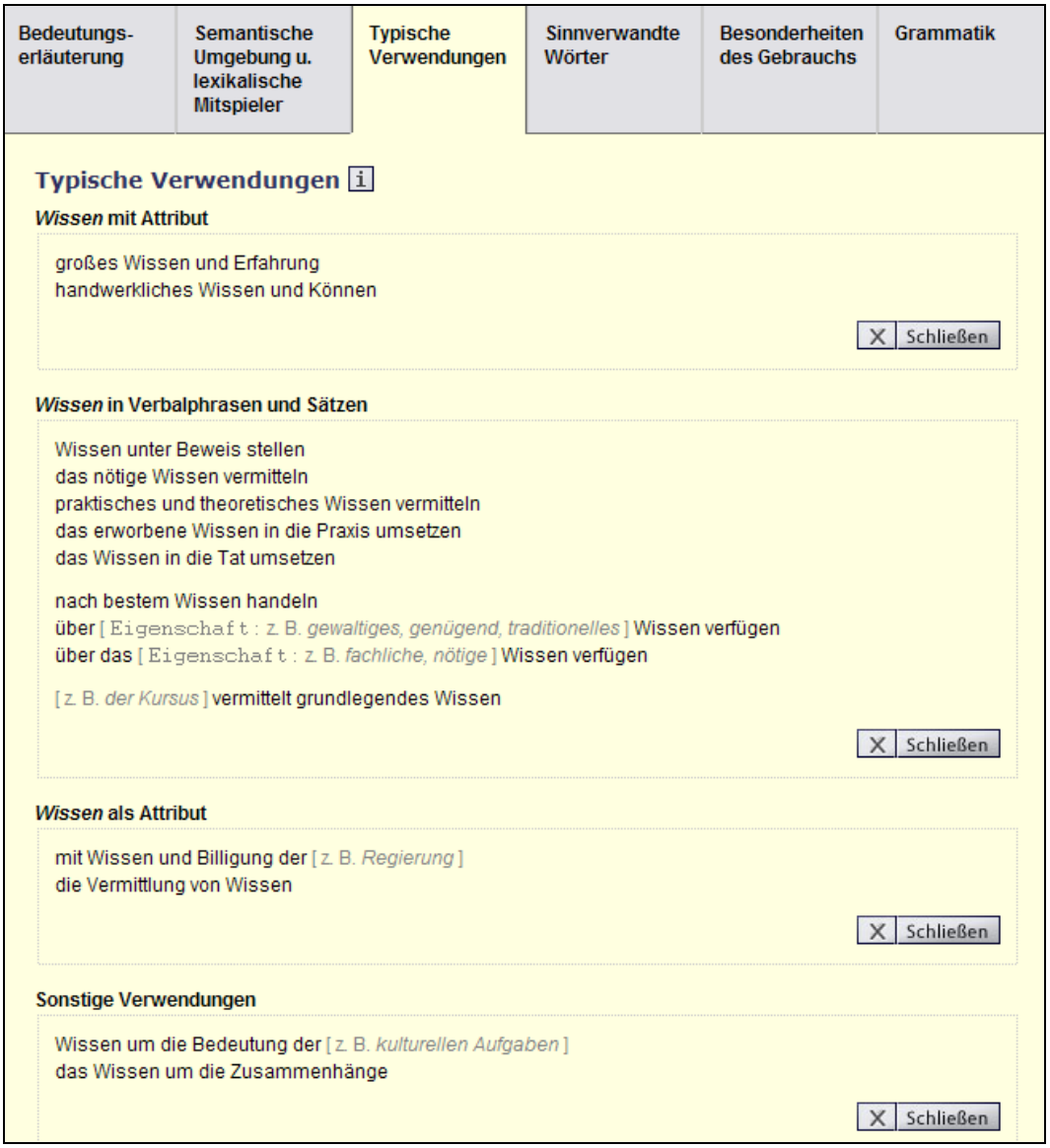

Figure 2: Patterns with *Wissen* in its sense 'Kenntnisse'.

Each major class is given a heading in order to allow the types of pattern to be identified quickly and to present the syntactic function of the lexical item in question. Within each major class, the syntagms are sorted according to specific criteria. For nouns, these include questions of whether the noun is typically used in specific frames. This includes aspects such as whether the head word is associated with a determiner or modifier, whether it is followed by a prepositional phrase, whether the noun functions as subject or object within the given phrase, etc. In some cases, there is a fourth major class where typical binominal coordinated structures or complements of nouns are listed. So, within the four main types of pattern, further subgrouping allows for a more specific pattern arrangement.

The sequences of *Wissen* in verbal phrases and sentences, for example, are divided into noun groups of which the head word is in the accusative, followed by infinitive structures. The second group contains prepositional phrases, followed by a group with *Wissen*, together with a finite verb preceding the head word.

Grammatical functions also play an essential role in classifying colligational patterns of adjectives. In example 2, the adjective *überflüssig* (*redundant, superfluous*) is presented in its sense 'entbehrlich' (dispensible).

Figure 3: Patterns with adjective *überflüssig* in its sense ‘entbehrlich’.

The association of the head word with nouns, verbs or adjectives is the determining factor in the classification system of adjectival head words. These are categorised according to whether the adjective displays an attributive, predicative or adverbial function. Internal criteria also take into account whether there are prepositions, or whether the adjective colligates with finite or infinitive forms of a verb, and so on. For instance, within the first category, the phrases are arranged into a group with the combination adjective and noun, a group with preposition, adjective and noun, and a group containing a verbal phrase with adjective and noun.

Verbs (not shown here) are presented according to structures where they occur in their finite or infinitive form and in patterns where the verb is used as an adjective in its participle form. Within these major head groups, further internal sorting considers aspects of occurrence within prepositional phrases, tense forms, the distinction between active and passive constructions, and combinations with or without modal verbs. Internal arrangement of the third

group, where the verb is used as an adjective, reflects its predicative, attributive or adverbial usage.

3.2.2 Alternative options: semantic slots, categories and illustrative fillers

The two examples in figures 1 and 2 also show that some patterns contain a bracket element. These are components within a construction which have an optional element which is not filled with one statistically significant or typical lexical item but which often has various lexical realisations, a so-called flexible, variable lexical filler. In such cases, the bracket element exemplifies a further element as part of the lexical sequence placed at a specific position. This is represented by exactly one lexical item which often belongs to a semantic-referential category. Prototypically, these are syntactically filled by subjects, objects, attributes, etc. From a semantic-conceptual point of view, these categories can be a subject which refers to a person, an adjective which often denotes a characteristic or state, or an illustrative verb designating a process or action. In some cases, the category is given a label such as [`Eigenschaft`] (characteristic), [`Person`] (person), [`Name`] (name), [`Ort`] (place), [`Handlung`] (process), [`Zahl`] (number) etc. In other cases, the lexical realisations vary so strongly that no common concept can be assigned to them, as is often the case with object slots. In these instances, no semantic category is provided and the bracket element is filled with one or two exemplifying fillers. These are chosen exclusively from the underlying corpus to illustrate authentic language examples.

For the lexicographic documentation, as was seen in figure 1, the filler is then taken in the corresponding grammatical form which is required by the pattern, and not in its basic form, such as the nominative case for nouns or the infinitive for verbs. In this way, syntactic features of the immediate lexical context are incorporated. Its illustrative status is pointed out by the introductory “z. B.” (e.g.).

In order to distinguish the bracket element, its semantic pattern and its actual lexical realisations, different typographical styles were chosen for their lexicographic presentation. Both different colours and different font types help to identify the necessary concept behind the integrated element and their illustrative status as lexical realisations. The brackets which remain empty stand for a whole clause, usually for following subordinate clauses. Tests are currently being carried out to determine whether users will have a better understanding of how to use the information given and what type of semantic and syntactic content belongs to the entire combination.

3.2.3 Usage notes

It has been observed that a number of patterns exhibit particularities with respect to specific contextual features. Therefore, another feature that has been included are buttons which open a separate box with additional usage notes. These explain specific semantic or syntactic restrictions or pinpoint observations with respect to the use of the word in question, as made in the corpus. They can refer to a single pattern, a specific group or an entire class of combinations.

Restrictions in terms of regional distribution are documented, for instance when a pattern is typically used in Austrian or Swiss German. Another case where an additional lexicographic note is provided is when a noun in patterned sequences is typically used in its plural form, e.g. *Bedürfnis* (*needs*) in its sense ‘Wunsch’ (wish) (see figure 4).

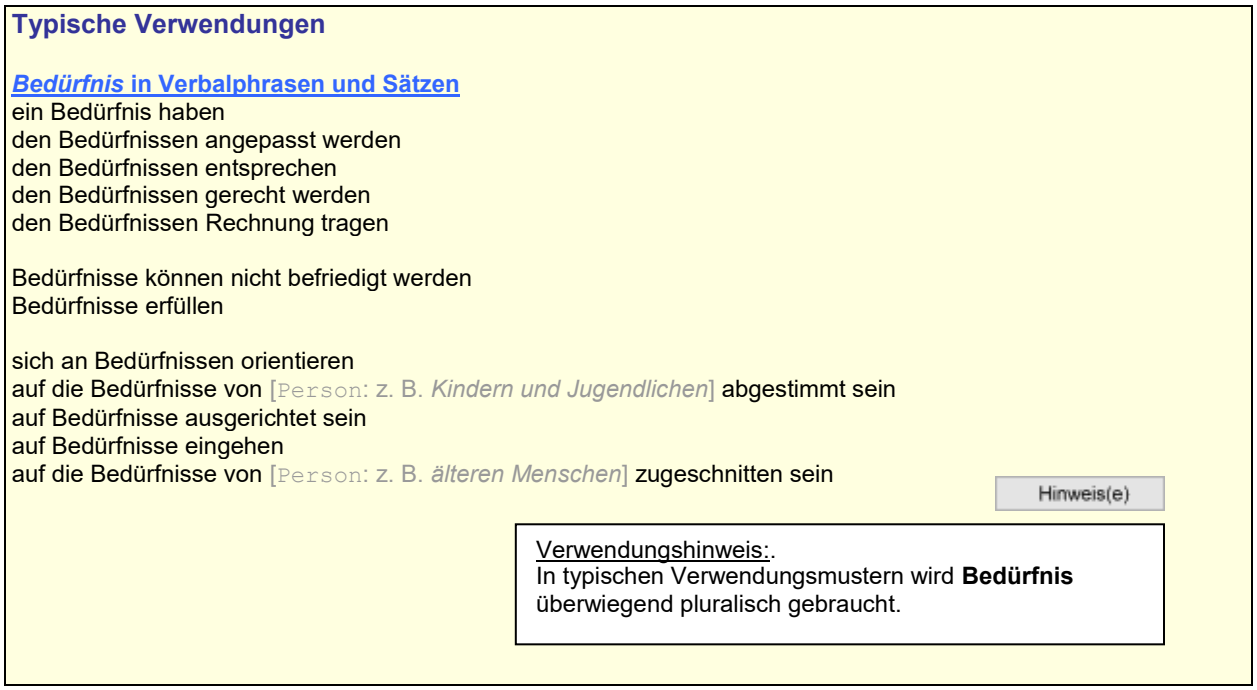

Figure 4: Usage note for patterns with *Bedürfnis* in its sense 'Wunsch'.

Alternatively, some adjectives tend to be used in their comparative form (e.g. *billig* (*cheap*) in its sense 'kostengünstig' (reasonable) (see figure 5)).

Figure 5: Usage note for patterns with *billig* in its sense 'kostengünstig'.

Certain other observations, for instance that a word such as *Frau* (*wife*) typically colligates with possessive pronouns when used in the sense of 'verheiratete weibliche Person' (married female person), are also included in usage notes (see figure 6).

Figure 6: Usage note for patterns with *Frau* in its sense 'verheiratete weibliche Person'.

It is also pointed out, for example, that a head word such as *mehr* (*more*) in its sense 'nicht weiterhin' (not any longer) typically colligates with negatives in more complex constructions (see figure 7).

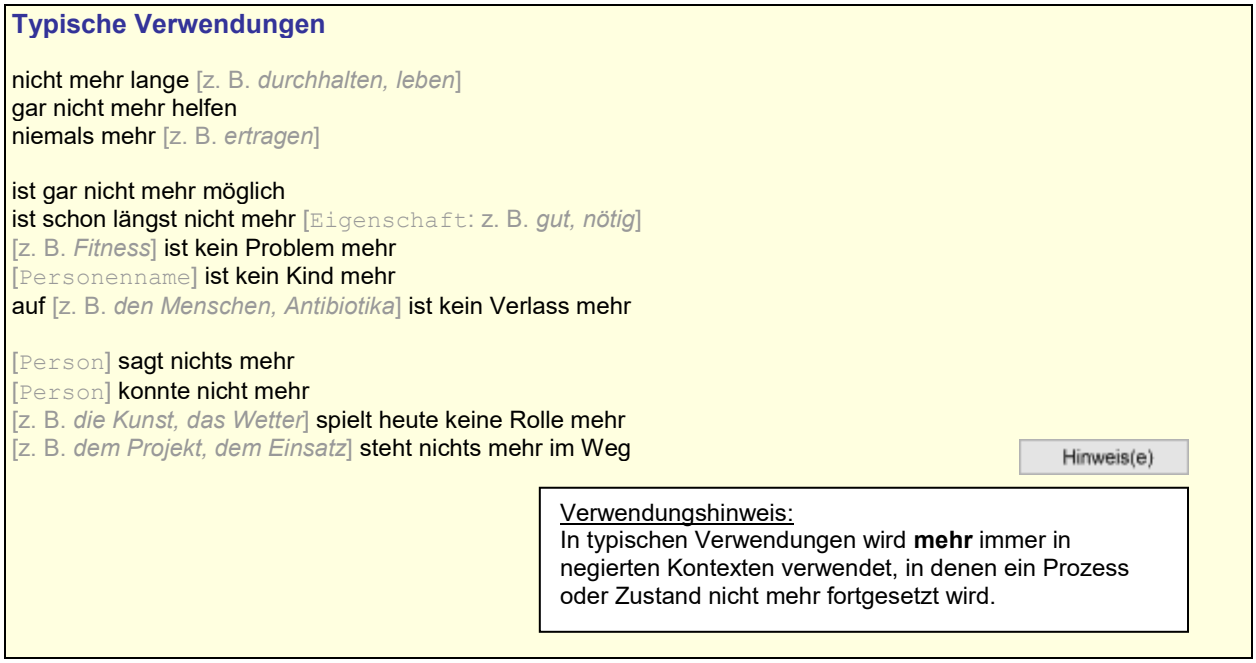

Figure 7: Usage note for patterns with *mehr* in its sense 'nicht weiterhin'.

In contrast to the two dictionary entries discussed in section 3.1 where no usage notes are provided, ELEXIKO tries to combine grammatical notes together with their lexical strings as they are rooted in some characteristic lexis.

Traditionally, indications of usage in printed dictionaries tended to be done through labels, specific coding or abbreviations. However, lack of space has prevented commenting in traditional reference works on the particularities of larger patterns. As there are no space restrictions in an electronic dictionary, complementary usage notes are provided where necessary, and they are composed in prose style and in full sentences to provide explicit explanations of the observations and co- and contextual constraints.

3.2.3 Looking back

The current presentation of typical and more complex syntagmatic patterns was introduced in 2007. It is the result of a fundamental formal revision of this particular kind of dictionary information. This revision took place for a number of reasons. One the one hand, ELEXIKO strove to implement insights about natural language, about words in patterns and constructions in its lexicographic descriptions. On the other, no German dictionary provided convincing guiding principles on how to document larger, typical patterns systematically so that both semantically and syntactically similar structures could be illustrated together. Guidance and inspiration for the new practice came in particular from English reference works such as the "patterns grammars" by COBUILD and corpus-driven studies of the lexis, culminating in theories of the lexicon such as Sinclair's idiom principle (1991), Hunston & Francis's pattern grammar (2000) and Hoey's lexical priming model (2005). They share an interest not only in corpora, patterns and the interface between lexis and grammar but also in the implications of linguistic thought for more practical endeavours and the connection between theory and lexicography, as well as teaching and learning a second language.

Initially, all the patterns of a lexeme were simply arranged according to complexity, an arrangement which had been temporary from the start of the project. It should be pointed out that ELEXIKO is a reference work which is being compiled from scratch and that it was only launched online in 2003 with about 250 illustrative entries. After three years of writing lexicographic entries, about 500 complex dictionary entries were available for an investigation of what typical colligation patterns occur in the corpus for what kind of head words. This means that the systematisation of this temporary arrangement of patterns remained an on-going task for several years until lexicographers could evaluate an appropriate amount of data. Until then, typical patterns of a head word were listed as illustrated in figure 8.

Typische Verwendungen

[…] vermittelt grundlegendes Wissen
das nötige Wissen vermitteln
die Vermittlung von Wissen
großes Wissen und Erfahrung
handwerkliches Wissen und Können
nach bestem Wissen handeln
über […] Wissen verfügen
Wissen unter Beweis stellen
praktisches und theoretisches Wissen vermittlen
das Wissen in die Tat umsetzen
Wissen um die Bedeutung der […]
das erworbene Wissen in die Praxis umsetzen
mit Wissen und Billigung der […]

Figure 8: Old presentation of typical syntagmatic sequences

With the help of the existing ELEXIKO-dictionary entries, a systematic analysis was carried out to identify the characteristics of larger syntagms. The result of the investigation led to the development of classification principles for each word class which reflect typical types of usage of the head entry within regular syntagmatic strings. The principal objective of the new systematisation was to group together patterns with similar functions and properties, an aspect that had to be ignored temporarily with the old arrangement. Another task was to find a more user-friendly solution to the bracket element which thus far had not provided any insights into the potential conceptual-semantic, lexical and grammatical realisations of the relevant item.

For adverbs, finding appropriate classification principles has so far not been part of the overall, larger classification model. This is a task that remains to be undertaken in the near future, as soon as there are enough adverbial head words which are fully lexicographically described and enough material has been built up for a closer examination.

At this stage of writing the dictionary, such fundamental changes to the presentation of the typical patterns of a head word were still possible without major disruption to the on-going compilation process. As the dictionary is an online resource with an XML-architecture, the underlying data-model (document type definition) needed adjustments to reflect new classes, types and groups of patterns. As a consequence, all existing entries needed re-editing, and real slot elements and their semantic categories needed to be added. Changes to the XML-structure further required adjustments to the XLST-style sheet which transforms data to an HTML browser view.

3.2.4 Data retrieval

As pointed out before, any lexicographic information in ELEXIKO is based on a comprehensive corpus.[2] For the detection of patterned frames, this implies that issues of co-selection are regularly brought up by corpus evidence. In particular, the analysis of co-occurrences and concordance lines brings together the disciplines of lexicology and grammar. As Partington notes:

> Concordances of semantically similar lexical items […] can be studied, and students will inevitably discover differences in use which are not contained in grammars and dictionaries. (Partington 1998: 47)

The underlying corpus was specifically built for the lexicographic purposes of the ELEXIKO project, and it is explored systematically with the help of its query tool COSMAS II[3], a software suite with its collocation and concordancing software "Kollokationsanalyse und Clustering"[4]. The corpus contains about 1.5 billion words of running text, mostly newspaper texts. By performing a collocation analysis, the lexicographer obtains statistically significant surrounding lexical partners for the search word and, additionally, more complex structures in which these partners are embedded with the search word. This is the basic methodology which is being used to detect actual lexical realisations for variable elements in prototypical syntagmatic patterns, a corpus-driven procedure for the analysis of language where data is approached without prior expectations. This means that the corpus is not used to search for examples the lexicographer has in mind or to validate his or her intuitive assumptions.

Traditionally, corpus-driven analyses of significant syntagmatic constructions have not been conducted for lexicographic purposes, although corpora and their tools offer quick and systematic access to colligational structures. As Tognini-Bonelli (2001: 26) emphasises, grammatical and lexical patterns can be observed at the same time, which is a huge bonus for lexicographers and for the description and documentation of usage patterns with lexico-syntactic characteristics in particular. Not only does the tool that is used to search the underlying elexiko-corpus provide an analysis of collocation but it is also able to detect syntagmatic patterns of the head word together with its collocates. Figure 9 is an extract from the results of a collocation analysis of the search word *Wissen*.

Total	Anzahl	LLR	Kookkurrenzen	syntagmatische Muster
2508	11	3066	**vermitteln** praktisches	72% praktisches [und theoretisches] Wissen vermitteln
3154	634		vermitteln	82% Wissen [... zu] vermitteln
3215	30		**theoretisches** Praxis	100% ihr theoretisches Wissen in die|der Praxis
3439	725	2795	**Erfahrung** grosses	70% grosses [...] Wissen und ... Erfahrung
5165	155	1887	fundiertes	98% ein fundiertes [...] Wissen
5169	144	1407	**erworbenes** Praxis umzusetzen	100% erworbenes Wissen in die Praxis umzusetzen
5585	273	1298	**Vermittlung**	64% der|die Vermittlung von Wissen
6323	98	951	**aneignen**	95% sich|das Wissen [...] aneignen
6460	126	857	**vertiefen**	89% das|ihr Wissen [... zu] vertiefen
6858	118	756	**umfassendes**	98% ein|sein umfassendes [...] Wissen
6977	26	731	**praktisches**	92% und ihr praktisches Wissen in
7229	155	638	**Fähigkeiten**	61% Wissen [und die ...] Fähigkeiten
7594	132	560	**erweitern**	87% ihr|sein Wissen [... zu] erweitern
7780	126	545	**Zusammenhänge**	88% das Wissen um|über die ... Zusammenhänge
7783	11	541	**Erfahrungen**	66% Wissen ... Erfahrungen
8381	56	522	**theoretische** umzusetzen	100% theoretische Wissen in die Praxis umzusetzen
8477	91	512	breites verfügt	85% Sie verfügt ... über ein breites Wissen
8788	218	509	**Bildung**	48% Wissen [und] Bildung
8842	51	501	**vorhandenes**	94% vorhandenes [...] Wissen
8926	78	486	**umfangreiches**	98% ein umfangreiches [...] Wissen über ...
9463	46	414	**Fertigkeiten**	65% Wissen und die Fertigkeiten
9532	69	406	**Geschicklichkeit**	57% Geschicklichkeit [und] Wissen
9620	88	393	**Weisheit**	55% Wissen [und] Weisheit
9668	88	389	**gesichertes**	100% gesichertes [...] Wissen
11188	32	249	**Nichtwissen**	81% das|zwischen Wissen [und] Nichtwissen
11285	55	246	**grosses**	98% ein grosses [...] Wissen
11486	76	239	**Kompetenz**	81% Wissen [und die] Kompetenz
11575	89	238	**verfügen**	70% über ... Wissen [...] verfügen
11614	19	230	**enzyklopädisches**	89% enzyklopädisches [...] Wissen
11717	72	227	**testen**	84% ihr Wissen [...] testen
11952	52	222	**Kenntnisse**	57% Wissen [und ...] Kenntnisse
12002	50	221	**Billigung**	94% mit Wissen [und] Billigung der|des ...
12259	41	209	**mangelndes**	100% mangelndes [...] Wissen über ...
12658	49	200	einbringen	97% sein|ihr Wissen [...] einbringen
14476	71	140	**Beweis**	98% ihr Wissen ... unter Beweis stellen
15368	27	111	**veraltet**	92% Wissen [...] veraltet
15382	14	109	**geballtes**	92% geballtes [...] Wissen
15405	23	109	**Verstehen**	78% Wissen und ... Verstehen

Figure 9: Extract of collocates and syntagmatic patterns of *Wissen*

The result is a list of collocates together with their common lexico-syntactic environment, and these are arranged according to their statistical status. The collocates are listed on the left-hand side with their significance value, and on the right-hand side the co-textual frames are given which are consistent in the corpus for the collocates associated with the search item. The marked constructions demonstrate some of the syntagmatic sequences as listed in the dictionary. The software identifies prototypical recurrent structures, and lexical and grammatical regularities can be identified in the co-text of the node word *Wissen*. The retrieved information is then analysed and lexicographically interpreted or generalised and, if relevant, assigned to a specific sense of the lexeme under scrutiny.

Not all patterns on the right in figure 9 are actually taken into consideration for the dictionary section “Typische Verwendungen”. First of all, not every pattern demonstrates a complex string of lexical items where lexical and syntactic relations are obvious. For example, simple combinations of adjective and noun such as *geballt* + *Wissen*, *veraltet* + *Wissen* and *umfassend* + *Wissen* are loose binary collocations. The same holds true for combinations such as noun + verb without prepositions (e.g. *Wissen vertiefen* and *sich Wissen aneignen*). These contain important semantic information, but they are listed in a separate section for semantic partners and their contexts.

Slot choices, where variable lexical items take a specific position within a construction, are also not filled with intuitive items. Every lexical choice is an example from a real discourse situation in the corpus. For example, in the case of *Wissen*, one significant collocate was *verfügen*, as in the construction *über […] Wissen verfügen* (*to have […] knowledge*). Here, the bracket element stands for various optional elements which are contextually realised by

adjectival elements. In such cases, it is concordances which help to scrutinise specific phrases. Concordances can easily be summoned from a corpus by a concordance programme. These lines of texts have the item being studied highlighted in the centre so that the immediate surrounding co-text becomes visible (see figure 10).

```
A98 im Dorf», die auch über das entsprechende Wissen aus der Geschichte verfügen,
A99   Gesundheitszustands der Fische über ein Wissen verfügen, das Behörden und
A99              in Gossau alle über dasselbe Wissen verfügen», erklärte er und fügte
A00           noch kaum über wirtschaftliches Wissen zu verfügen, die gleiche Frage.
V00    ausreichende Mittel und entsprechendes Wissen verfügen, erklärte Sickl.
E96   verfügen über wertvolles traditionelles Wissen.
E97     nur verlangt, dass sie über ein gutes Wissen verfügen und methodisch und
E97      zeigt das doch, dass Diebe über viel Wissen und eine gute Infrastruktur
E97       verfügen oft über ein unschätzbares Wissen zur lokalen Biodiversität.
E97  Südens müssten heute jedoch über eigenes Wissen und eigene Kapazitäten verfügen.
E98   einige Fragen, die vor dem historischen Wissen, über das wir heute verfügen, in
E00   die Redaktionen noch über das fachliche Wissen, über Zeit und Musse verfügen,
I99    über völlig unzureichendes politisches Wissen verfügen.
N93       Es gibt Künstler, die über geheimes Wissen verfügen.
N95     Man werde bald über bisher ungeahntes Wissen verfügen.
N96     werde es hingegen sein, über genügend Wissen zu verfügen.
N99  muss einer über allerlei psychologisches Wissen verfügen.
P96       an jene Leser, die über kompetentes Wissen verfügen.
R97  und Schülerinnen sollen über ein solides Wissen verfügen, aber sie sollen auch
R97      Fachleute verfügen über ein sicheres Wissen, wie sie helfen können".
R98    in drei Jahren über weltweit begehrtes Wissen verfügen.
R98     Wir verfügen über ein weit gestreutes Wissen zu einzelnen Fragen, die im
O97  sehr viele Experten, die über gewaltiges Wissen verfügen.
O99    ein beträchtliches zeitgeschichtliches Wissen zu verfügen scheint.
K99   deswegen verfügen sie über ein tieferes Wissen vom Mensch- Sein, von Liebe und
```

Figure 10: Extract from concordances for the lexical unit *über Wissen verfügen*.

A rapid scan of the concordances enables the linguist to detect actual variants within this specific construction. These are adjectival attributes such as *genügend* (*sufficient*), *gut* (*good*), *historisch* (*historical*), *wirtschaftlich (economic)*, *traditionell (traditional)* and so on, colligating with the node word, and they specify the corresponding noun differently. The lexicographer then chooses up to three illustrative lexical fillers from such concordance lines.

4 Lexicological demands and their lexicographic consequences

From a lexicographic point of view, it has so far been argued that the description of typical usage patterns which illustrate colligation is as necessary for the decoding and text production purposes of non-native speakers as collocation and semantic association. In addition, a documentation of typical syntagmatic sequences enables native speakers to capture the interface between lexical and grammatical characteristics and relations within larger constructions. And from a theoretical cognitive perspective, the examination of such patterns provides insights into how lexical sequences are construed. Hunston & Francis (2000) remark that the fields of lexicography, language education, second-language acquisition and linguistic theory share or should share an interest in lexical phrases. The following aspects in particular are pointed out:

> the frequency and therefore the importance of lexical patterns, the varying degrees to which lexical phrases are open to variation in wording, the functions of lexical phrases and the importance of lexical phrases to a model of language that gives lexis and grammar equal priority. (Hunston & Francis 2000: 8)

The question that remains to be answered is how these aspects are reflected in the work of ELEXIKO. In a conceptual phase, the lexicographic presentation of complex syntagms has been the centre of attention and these aspects were taken into account when new reflections on this part of the dictionary led to a complete rearrangement in 2007 (cf. Storjohann et al. 2007). Its main objective was to incorporate new theoretical insights and to guide the perception of users to the nature of colligational constructions as attested in real communicative situations. Some

lexicological ideas on frames and the idiom principle, as proposed by Sinclair (1991) and Sinclair & Renouf (1991), the theory of pattern grammar as put forward by Hunston & Francis (2000), and the notion of lexical priming as developed by Hoey (2005) have all become central points of departure for the lexicographic realisation of these aspects in the dictionary. As a common feature, their work is concerned with language patterns and approaches co-selection from different but often complementary perspectives.

Firstly, the corpus-driven methodology as applied in ELEXIKO to retrieve colligation structures ensures that lexicographic documentation of usage patterns is based on attestation of frequency and distribution, and it implies statistical justification. Comprehensive corpora and their tools allow for the exploration of recurrent sequences of items which otherwise could only be identified though introspection. As the lexicographic reality shows, a large number of patterns are simply not recorded in reference books. In addition, a lexicographer cannot make judgements on the frequency and statistical significance of lexical cohesion. With corpora, lexical choices can be studied empirically and with the help of collocations and concordance lines, patterns emerge. As lexicographic practice using corpora continuously shows, some findings and observations are unexpected. This, of course, questions the reliability of human intuition (cf. Tognini-Bonelli 2001: 86). However, on the other hand, introspection and a good lexicographic competency are indispensible for the analysis and interpretation of the retrieved language data.

Secondly, corpus data best reveals to what extent lexical routines contain optional variation and how these can be potentially realised in text and discourse. The need to present variation has been implemented by exemplifying illustrations of lexical realisations in their required semantic and grammatical forms.

Thirdly, the lexical phrases are partly categorised, and their corresponding headings summarise specific textual functions of the reported patterns. Syntagmatic prefabricated strings performing similar textual and propositional functions are presented as corresponding discourse types and their subgroups. It is through this categorisation that it is possible to documente patterns together.

Finally, the fifth aspect that an analysis of lexical phrases must account for in a model of language where lexis and grammar are not two separate domains is an aspect that is discussed and largely agreed upon now in linguistic theory. Lexicographic analysis of corpus data confirms that lexical and grammatical features are blurred in larger sequential patterns and any adequate description – irrespective of its aims should account for both characteristics.

4.1 Pedagogic objectives

So far, it has been emphasised that the main advantage of the presentation of colligational patterns is to provide detailed insights into the construction of lexical sequences and to show typical, rather than possible, ways of building more complex structures. ELEXIKO's underlying principle, to document regular and not random occurrences of patterns as found in the corpus, cannot account for some learners' needs to obtain information on what is not possible or what is explicitly incorrect. Just like any other data collection, the compilation of a corpus needs to comply with certain linguistic criteria. Nonetheless, a closer look at any comprehensive corpus always reveals the diversity, flexibility and unexpectedness of language, as well as its potential to change. This poses difficulties for lexicographers when providing information about impossible patterns. However, by spelling out what the typical associations of a word are, an entry "offers an extremely valuable shortcut to a lexical item's characteristic collocations and colligation" (Hoey 2005: 186).

The organisation of typical colligational patterns offered in ELEXIKO and the number of such patterns made available in it should also be considered to be a resource for learners and teachers. Targeted searches allow for the retrieval of words which behave in similar ways and

recur in similar patterns, e.g. nouns which are followed by the same preposition, or verbs which tend to be used primarily in their participle form, etc. Combined searches for synonyms, for example, allow for direct comparison of patterns. Alternatively, for instance, in the case of searches for nouns which are often used in their plural form and together with adjectival attributes, users can obtain lists of words with similar colligational behaviour. Groups of words sharing the same kind of noun or verbal fillers can be identified, or groups of verbs denoting speech acts that are followed by a *dass*-clause (that-clause) can be collected. Generally, words with similarities or differences in lexical co-textual usage become readily apparent. Through this, teachers are able to collect material for exercises in the classroom and learners are able to extend their vocabulary within a specific notional topic. And as Hunston & Francis (2000: 271) note, "learning strings rather than individual words enable the learner to compose lengthy utterances with the minimum of effort". Ideally, the reference book is designed not only to provide new information but also to raise awareness of existing knowledge and to raise implicit information to a more conscious level.

5 Conclusion

To sum up, the main benefits that this part of the ELEXIKO-dictionary can offer are a systematic classification of types of pattern which, above all, illustrate the function of the head word in smaller or larger constructions. These functions are explicitly spelled out by headings. Patterns with similar semantic and syntactic features are grouped together and hence also demonstrate related combinations. The colligational structures are authentic examples of language use because they are taken from a corpus made up of concrete instances of language use. All patterns which are listed can be considered to be typical and regular. They are not single or random occurrences of use. Variable and optional elements are accommodated and, finally, additional usage notes complement given information on specific structures. All these features allow for better language guidance for learners of German. Of course, one has to admit that generally the amount of information presented in ELEXIKO cannot be offered in a printed version of a single volume reference book. However, it is not just the amount of information in traditional dictionaries which is subject to criticism.

Undoubtedly, the use of corpora and elaborated computational assistance to retrieve empirical quantitative evidence of colligation structures also has its limitations, but as Stubbs (2009: 23) points out, they are "better than trying to discover patterns by introspection". The drawbacks of introspective explanations can be summarised as follows:

> Introspective data underestimates the complexity and subtlety of speech acts: this is shown by the substantial empirical data which we now have for investigating what speakers do with words. (Stubbs 2009: 29)

However, seeking to account for what speakers actually do with words means more than just incorporating corpora and their tools as basic working materials and methods into lexicography. Corpus material should not be brought in as an "extra bonus" but rather as a "determining factor" (Tognini-Bonelli 2001: 66). It is corpus-driven and not corpus-based procedures which can provide systematic methods for collecting relevant data and information on recurrent and typical, instead of random and possible, phrases and patterns.[5] It is this methodology which allows findings to be derived inductively from language material. This is in keeping with Tognini-Bonelli's view (2001: 110) that

> it is hoped that more studies will systematically show interrelations and the overlaps […] and these will be reflected and reported accordingly in the reference works […].

With respect to German lexicography, this means that the call for observation-based, data-drawn and theory-driven approaches needs to be answered, in place simply of corpus validation and exemplification need to be answered.

Notes

1 ELEXIKO is publically accessible without registration at www.elexiko.de. It is an integrated reference work under the roof of the lexicograhic portal OWID also accessible at http://www.owid.de/elexiko_/index.html.

2 For further information on the ELEXIKO-corpus, see Storjohann (2005).

3 COSMAS = *Corpus Search, Management and Analysis System*. See also http://www.ids-mannheim.de/cosmas2/.

4 See project "Methoden der Korpusanalyse und -erschließung" at http://www.ids-mannheim.de/kl/projekte/methoden/. The tool "Statistische Kollokationsanalyse und Clustering" is based on statistical evaluation of mass data and was developed by Cyril Belica (1995-2002).

5 For further explanations of the differences between corpus-based and corpus-driven investigations, see Tognini-Bonelli (2001).

References

Belica, Cyril (1995). *Statistische Kollokationsanalyse und Clustering. COSMAS-Korpusanalysemodul*. Mannheim: Institut für Deutsche Sprache.

Busse, Ulrich and Anne Schröder (2009). "Fowler's Modern English Usage". In Ute Römer and Rainer Schulze (eds) *Exploring the Lexis-Grammar Interface*. Amsterdam and Philadelphia: John Benjamins, 69-87.

Hanks, Patrick (1987). "Definitions and explanations". In John Sinclair (ed.) *Looking Up – An account of the Cobuild project in lexical computing*. London and Glasgow: HarperCollins. 116-136.

Haß, Ulrike (2005). "Elexiko - das Projekt". In Ulrike Haß (ed.) *Grundfragen der elektronischen Lexiographie. Elexiko – das Online Informatiosnsystem zum deutschen Wortschatz*. Berlin and New York: de Gruyter, 1-17.

Hoey, Michael (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Hoey, Michael (2009). "Corpus-driven approaches to grammar". In Ute Römer and Rainer Schulze (eds) *Exploring the Lexis-Grammar Interface*. Amsterdam/Philadelphia: John Benjamins, 33-47.

Hunston, Susan & Francis Gill (2000). *Pattern Grammar. A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam and Philadelphia: John Benjamins.

Sinclair, John and Antoinette Renouf (1991). "Collocational frameworks in English". In Karin Aijmer and Bengt Altenberg (eds), *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman, 128-144.

Sinclair, John (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, John (1996). "The search for units of meaning". *Textus* 9, 75-106.

Storjohann, Petra (2005). "Typische Verwendungen". In Ulrike Haß (ed.) *Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz*. Berlin and New York: de Gruyter, 235-248.

Storjohann, Petra/Möhrs, Christine (2007). "Das elexiko-Wörterbuch im Wandel - Typische Verwendungen neu beleuchtet". In *Sprachreport* 3/2007, 12-16.

Storjohann, Petra (2005). "Das elexiko-Korpus. Aufbau und Zusammensetzung". In Ulrike Haß (ed.) *Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz*. Berlin and New York: de Gruyter, 55-70.

Stubbs, Michael (2001). *Words and Phrases. Corpus Studies of Lexical Semantics*. Oxford: Oxford University Press.

Stubbs, Michael (2009). "Technolgy and Phraseology". In Ute Römer and Rainer Schulze (eds) *Exploring the Lexis-Grammar Interface*. Amsterdam and Philadelphia: John Benjamins, 15-31.
Teubert, Wolfgang (1998). "Korpus und Neologie". In Wolfgang Teubert (ed.) *Neologie und Korpus*. Tübingen: Gunter Narr, 129-170.

Tognini-Bonelli, Elena (2001). *Corpus Linguistics at Work*. Amsterdam and Philadelphia: John Benjamins.

English Reference Works

COBUILD (1987). *Collins Cobuild English Language Dictionary*. London: HarperCollins.

Francis, Gill, Susan Hunston and Elizabeth Manning (1996). *Collins Cobuild Grammar Patterns 1: Verbs*. London: HarperCollins.

Francis, Gill, Susan Hunston and Elizabeth Manning (1998). *Collins Cobuild Grammar Patterns 2: Nouns and Adjectives*. London: HarperCollins.

German Reference Works

ELEXIKO: www.elexiko.de

LANGENSCHEIDTS *Großwörterbuch Deutsch als Fremdsprache* (1999). Götz, Dieter (ed.). Berlin u. a.: Langenscheidt.

WAHRIG *Großwörterbuch Deutsch als Fremdsprache* (2008). 1st edition. Gütersloh and Berlin: Wissen Media Verlag.